Madroño, Vol. 47, No. 4, pp. 265–268, 2000

ELECTRONIC ACTIVITIES OF THE UNIVERSITY AND JEPSON HERBARIA

Richard Moe
University and Jepson Herbaria, 1001 Valley Life Sciences Building,
University of California, Berkeley, CA 94720-2465

This review treats computer activities that are carried out as part of the institutional agenda of the University and Jepson Herbaria. Of course, individual workers in the Herbaria depend on a variety of programs daily: e-mail has replaced to some degree letters and telephone calls and has gradually claimed an increasing part of our time. Word processors and spreadsheets have nearly completely replaced typewriters. Programs that manipulate DNA sequences and generate hypothetical phylogenies are used by several of the staff and their students. Institutional use has as its nucleus the database developed by the Specimen Management System for California Herbaria (SMASCH), the continuation and maintenance of *The Jepson Manual,* the revival and furtherance of Jepson's Flora, publication of information dealing with nomenclature (including typification), and publication of information dealing with the history of the Herbaria.

The electronic activities can be grouped into three broad classes (which overlap considerably):

Curation—concerns of day-to-day operations of the Herbaria;

Publication/Education—information made available to colleagues and the general public

Interactive outreach—uses which allow the Herbaria to benefit from the knowledge of the general public, as well as outside botanists, both amateur and professional.

History

Computerization of the Herbaria began in 1982 when Thomas Duncan, a pioneer in computerized taxonomy, assumed directorship. Supported by a succession of grants from the National Science Foundation, computerization was fostered by the Specimen Management System for California Herbaria (SMASCH) which began in 1991. As SMASCH got underway, Duncan moved to found the University Museum Informatics Project, which was closely associated with the development of SMASCH. Although SMASCH was conceived as a confederation that would include many western herbaria, only the University (UC) and Jepson Herbaria (JEPS) were able to obtain sufficient funding to proceed. SMASCH developed protocols for organizing the tremendous variety of specimen information into databases and for automating herbarium administrative procedures. The project consisted of a coordinator (Thomas J. Rosatti), a software specialist (Mickey Ellinger), and a data-entry crew, some of whom remained for nearly the whole project and others of whom were transient.

The SMASCH software comprises a Sybase relational database that is accessed by data-entry routines incorporating the X-window graphical user interface. The original goal of SMASCH was to capture all label and annotation information for each vascular plant specimen from California in UC and JEPS, and to supplement this with a high-resolution image. It became apparent early on that these goals were too ambitious, and as a result, imaging was restricted to specimens of special importance, and recording complete annotation history was abandoned.

By the end of its final funding period in 1999, SMASCH had computerized more than 300,000 specimens—all of the California accessions and all of the North American holotypes (which were among the specimens imaged). During the project, emphasis was placed on optimizing efficiency of retrospective data entry and as a result, optimal procedures for data retrieval were deferred. At present, data can be accessed via X-window screens, via direct Structured Query Language (SQL) queries against the Sybase data tables, or via a web interface. The first two methods are not usable directly by the public at large, but the web interface is available to everyone: http://www.mip.berkeley.edu/www_apps/smasch/. This interface, which is not yet complete, now allows users to produce a list of all specimens that:

1. have a certain scientific or common name, or
2. occur in a selected county (or counties), or
3. were collected by a given collector, or
4. were collected on a given date or range of dates, or
5. contain designated "voucher" information, e.g., indication of flower color, chromosome count, habitat information.

The criteria can be combined so that it is possible, for instance, to request specimen data of all specimens of *Pinus sabiniana* Douglas collected by Jepson between 1895 and 1900 in Napa or Solano counties. The interface also allows queries by collecting event (each collecting event being a unique combination of collector, date, and location). Thus one could retrieve all collecting events by Jepson in Amador County, and from that one could obtain any or all of the specimens collected at a given location. Web queries are made not against the

main database, but against tables extracted from the main database and optimized for retrieval speed. Therefore, the extract that is available on the web may lag behind the main database and not immediately reflect updates.

It is also possible to retrieve lists of names grouped by county for which the Herbaria have vouchers. This feature was added recently at the request of people constructing county-level floras and is an example of how we hope to modify our web publications to serve the needs of the botanical community.

In addition, the raw data have been made available to the CalFlora Database http://www.calflora.org/, where they constitute most (more than eighty percent according to figures on the CalFlora "Information about Datasets" page) of the specimen data available at the site.

The future of SMASCH. The database that was established by the SMASCH project will be maintained by the staff of the Herbaria and will be expanded as time and funding permit. The following database tasks are now part of Herbaria routine: Modification of tables to reflect revisions of identification or nomenclature; data entry of newly accessioned California material or of returned loans; corrections of inconsistencies in the data. The original methods of data entry were designed to capture information from sheets in the Herbaria—retrospective data capture. In the future we will have the chance to computerize specimens that are not yet accessioned, and for this we have explored new methods of data entry. Most collectors now make labels for specimens that they eventually deposit in UC/JEPS from databases or other computer files. When we receive new acquisitions in lots of several hundred, it works well to convert these databases or files to an intermediate format from which selections can be bulk loaded into the database. This reduces data entry to associating the collector's number with a barcoded accession number. This is being done now with new accessions from Dean Taylor, Lowell Ahart, and Vernon Oswald, as well as with several lots of specimens in our "backlog." We anticipate being able to handle the more than 100,000 bryophyte specimens deposited by Daniel Norris and specimens deposited by James Shevock similarly. We are also experimenting with data input via web forms. We have made available a label-printing form that can store data sent to it so that if the specimens in question are ultimately received, the corresponding data can be retrieved and transferred to the main database. We also have a web form that will allow curators to enter information from any previously unrecorded type specimens that they encounter in the main collection. We are working to enrich the web query interface to the specimen database in order to allow a greater range of queries and to permit users to provide feedback automatically keyed to the specimen or name they are dealing with. There may be users who are willing to help us rectify inconsistencies in the database, if we make the process convenient. For instance, a user might be able to add location information to a specimen by looking up another specimen collected by the same person on the same day.

Electronic products relating to The Jepson Manual. *The Jepson Manual* (Hickman 1993) is a widely used reference book that could be converted into an electronic product in numerous ways. Because the copyright is held by the University of California Press, however, use of the *Manual* other than in the form in which it was published has not been pursued. The electronic files from which camera-ready copy of the *Manual* was printed have been translated in part into Extensible Markup Language (XML), and from this version we have extracted the names of the taxa and a variety of associated data, including distribution. Distribution is indicated in the *Manual* by citing the bioregions in which a taxon occurs. Bioregions are hierarchical (Hickman 1993, pp. 37–48): e.g., the Great Central Valley (GV) comprises the Sacramento Valley (ScV) and the San Joaquin Valley (SnJV), and is itself part of the California Floristic Province (CaFP). We have made a web application (http://ucjeps.herb.berkeley.edu/jeps-list.html) that uses the base map of bioregions from the *Manual.* The distributions are displayed on the map by expanding the composite regions and coloring in each smallest unit. The distribution records are modified as new information becomes available, and since the maps are constructed on request, they reflect current understanding of ranges. Because of this dynamic generation of the maps, there are no static pages for outside sites to link to directly. Furthermore, the URL for the page of a given taxon incorporates a compressed representation of the distribution, so the URL changes when the distribution changes. To circumvent this problem we maintain a simplified, but slower, access procedure for external links, whereby an incoming request that includes just the taxon name can be associated with other information. The tremendous quantity of taxonomic information available on the web often makes easy what was previously difficult or impossible. Much of the value of the web arises from hypertext links, but links are not easy to maintain—URL's change, out-of-date URL's remain in caches of indexing sites, methods of generating dynamic pages change with software modification.

One of the challenges of website development is making pages flexible enough that they may be used in ways that the authors haven't anticipated—without causing security problems. It is important to make each page independent of the pages to which it is linked from the main site, because context may be lost when visitors come from an unintended page, or from an index cache. Therefore, we try to identify each page—whether it is dynam-

ically or statically generated—with a title that is meaningful in any context, and that has links back to the Herbaria home page.

The Jepson Interchange. We have for some time been considering the possibility of using web technology to simultaneously track changes in California floristics, make available expanded treatments of California plants (i.e., more extensive than those in *The Jepson Manual*), prepare for the second edition of *The Jepson Manual,* and communicate with amateur and professional botanists interested in the California flora (see http://ucjeps.berkeley.edu/far_west_initiative.html for an exposition by Barbara Ertter of a distributed information system for native and naturalized plants). This project is now underway, having been enabled by a grant from the William R. Hewlitt Revocable Trust. The new project, called the On-line Interchange for Advances in California Systematics, or the Jepson Interchange (http://ucjeps.berkeley.edu/jepson_flora_project.html), will provide a continuously updated authoritative list of California vascular plants, provide treatments for taxa not covered in *The Jepson Manual,* account for names not included in the list (synonyms, misidentifications), and most importantly, provide a convenient forum for all interested persons to assist in cataloging the California flora and understanding California plants. We now have web forms which can be used to report new records, call attention to publications pertinent to the flora, revise distributions, suggest hyperlinks, or contribute other information. The contents of the forms will be stored automatically in a database and e-mailed to an editor who will be the first stage of an authorization filter that may also include *Jepson Manual* editors and authors and other specialists. All submitted information will be available on the web, but only information authorized by the editorial committee will be incorporated into the list of taxa. As an adjunct to the Interchange, and in collaboration with the Digital Library Project of the University of California and Xerox PARC, we will make available scanned images from Jepson's "*A Flora of California,*" which contain a wealth of details on California plants.

Other Web Publications

Index Nominum Algarum (INA). The INA (http://ucjeps.herb.berkeley.edu/rlmoe/) is a card file maintained by Paul Silva at the Herbarium of the University of California. It contains nearly 200,000 names of algae (in the broad sense). Associated with the INA is a separate card file containing bibliographic references pertaining to algal taxonomy—the Bibliographia Phycologia Universalis, or BPU. Cards that have been added since 1988 have been printed from a database, and the data are available on the web. As a preliminary step towards entering the remaining cards (pre-1988) in a database, and to provide archival protection, the cards have been scanned as digital images. These digital images can be used in conjunction with indexes approximately like the physical cards can be used, but they are available from more than one site. We are making indexes to the images in two ways: via optical character recognition, and via forms that allow users to help by entering index entries directly.

Purpus site. Carl Albert Purpus was a plant collector in western North America with an unpaid curatorial appointment at Berkeley. Barbara Ertter and Tom Schweich have innovatively combined the wealth of archival material in the Herbaria with specimen information to present historical, floristic, and related data about the North American collections of Carl Albert Purpus in a globally accessible and informative manner for use by students, historians, botanists, and interested laypersons (http://ucjeps.berkeley.edu/Purpus/). We hope to have a variety of similar web publications in the future.

Indian Ocean Catalogue. The Indian Ocean Catalogue (Silva et al. 1996) is a compilation of all published records of species and infraspecific taxa of benthic marine algae from the Indian Ocean. Published by the University of California Publications in Botany in 1996, it was converted to a web version during the reviewing process (http://ucjeps.herb.berkeley.edu/rlmoe/tioc/ioctoc.html). The web version, which was generated by filtering some 75 files marked up for the troff typesetting program, allows a variety of searches, and is updatable by user input.

Conclusions

Because we are dedicated to increased computerization in the Herbaria, it is well to consider some of the assumptions and consequences of the process. The advantage of computerization is not that the traditional mission of the Herbaria can be accomplished more cheaply and more rapidly, although this is sometimes assumed. In fact, computerization probably makes routine tasks more time-consuming and cumbersome. The real advantage is that tasks can be accomplished that were previously impossible, with concomitant increase of the value of the specimens and specimen data to the Herbaria, to other institutions, and to the public. As a simple example, during the production of *The Jepson Manual* (Hickman 1993), it was not possible for contributors to check the vouchered distributions of species they were responsible for without borrowing all the specimens or visiting the Herbaria. When the second edition is assembled, it will be simple to provide all contributors with electronic reports of all the UC/JEPS specimens pertinent to their treatments, with distributionally noteworthy specimens flagged for their attention.

A specimen database is institutional. It, like the rest of the Herbaria, needs to function in perpetuity. Resources need to be allocated to it forever. It must transcend hardware, software, and personnel. A sig-

nificant cost of computerization is the requirement for vigilance—not with respect to privacy or security issues, though these are important—but with respect to changes in hardware, software, and personnel. All of the changes must be accommodated, and neither too rapidly, which would lead to constant turmoil, nor too slowly, which might cause intermittent large disruptions.

The Herbaria depend on the University for networking, database servers, and expertise. As our computer applications come to be used by outside users, those users will be similarly dependent. As Internet applications become more common, we likewise depend directly on outside institutions.

Literature Cited

Hickman, J. C. 1993. The Jepson manual: higher plants of California. University of California Press, Berkeley, CA.

Silva, P. C., P. W. Basson, and R. L. Moe. 1996. Catalogue of the benthic marine algae of the Indian Ocean. University of California Publications in Botany 79.